2022 31st IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 29 - September 2, 2022. Naples, Italy.

Group Activity Recognition in Restaurants to Address Underlying Needs: A Case Study

Ada V. Taylor[1], Roman Kaufman, Michael Huang, and Henny Admoni[1]

Abstract— **Enabling robots to identify when humans need assistance is key to being able to provide help that is both proactive and efficient. This challenge is particularly difficult for humans eating a meal in a restaurant, a context which is dense with interlaced social elements such as conversation in addition to functional tasks such as eating. We investigated the challenge of identifying human dining activities from single-viewpoint footage by collecting and annotating the individual activities of five two-person meals. From this process, we found that addressing the question of identifying meal phases and overall neediness requires identifying an underlying *group state* for the table as a whole. We report on the individual activities and group states, as well as the interdependencies between these factors that can be leveraged to both provide and measure effective robotic restaurant service. In addition to the insights revealed by this dataset, we describe preliminary attempts to create an automated classification system for these activities.**

I. Introduction

Restaurant activity recognition has the potential to be a valuable tool for automating goals of quality restaurant service such as monitoring guest satisfaction, identifying moments when assistance is needed, and assessing the flow of a meal to improve service [17], [19]. Providing this service requires identifying both the directly observable activities each guest is performing, as well as the underlying needs that are driving these actions and stages of meal progression.

Due to the nature of a meal at a restaurant, a naturalistic dataset of this kind is hard to capture. In laboratory environments, emotional cues are often different than in the wild [3]. Similarly, underlying motivators such as boredom and hunger need to exist in their natural state to properly model the experience of eating a meal in a restaurant and evoke the physical cues that are generated by these needs and resolved over the course of the meal. To address this, we collected publicly-available webcam footage from a real restaurant.

The tradeoff for this naturalism is that we cannot interview guests about their sentiments over time. We start by annotating the individual, directly observable actions of the guests to look for patterns of behavior. We detail the challenges involved in collecting and analyzing activities of this kind, as well as features of the restaurant environment specifically that can be leveraged for automated labeling of activities.

We characterize the activities found in this environment. We found that individual activities tend to take place in short bursts, with social activities interspersing longer tasks. We found that the activities of each guest can provide a useful indicator of the other guest's activity, particularly for social activities between guests. We outline how this relationship may be used to compensate for issues such as occlusion.

Fig. 1: *Individual table of interest, 450x400. The static elements of the restaurant enable automated detection of table members from their consistent locations, and their group membership from their presence at the table together. Joint locations are annotated with dots, with individuals A and B denoted by the bounding boxes around their chairs. The current label for both individuals' activities is "eating."*

Underlying guest motivations and needs may not be directly indicated through their observable actions, yet waiters make inferences about their levels of need in order to provide quality service. To identify the phases of service required of a robot waiter, we explore a collective *group state* that describes the overall status of guests sharing a meal. These group states reflect the activities guests are performing, and when service is required to continue the meal.

Despite group states representing an abstract concept, we provide empirical support for each group state having a distinctive profile of individual activities, and a consistent

*This work is supported in part by the National Science Foundation (#IIS1755823, #DGE1745016), the Sony Group Corporation, and Sony AI.

[1]Ada Taylor and Henny Admoni are with the Carnegie Mellon Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA adat@andrew.cmu.edu, hadmoni@andrew.cmu.edu

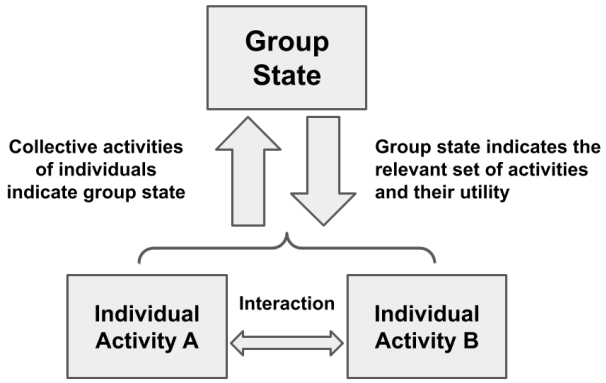

Fig. 2: *From our data, we propose that accounting for an underlying group state allows for better understanding the current needs of guests, and identification of when a transition in needs has occurred. This overall status is determined by the set of group members as a whole. We also observe that activities have different distributions based on the overall group state, as activities such as "eating" have different levels of focus over the course of a meal. We also see evidence for the activities of group members influencing each other, such as cycles of taking turns talking and listening.*

sequence over the course of a meal. We argue that the relationship between these activities and group state is strong enough that it can be leveraged to identify changes in the phase of service and therefore when assistance will be needed by guests. We contrast this approach with using a hidden Markov model (HMM) to infer the states underlying guest activities.

We conclude that labeling group states is required in order to teach a robot waiter when and what kind of assistance is needed by guests, particularly when the activities preceeding a change of state often look different depending on the overall group state. We also note that some group states have strong correlations with particular activities, like using tools such as a wallet, glasses, or purse when paying the bill. We propose that a shift to a larger proportion of social and idling behaviors may be a useful signal across many group states indicating that guests are ready for waiter assistance.

To illustrate these relationships which are key to providing effective restaurant service and measuring guest satisfaction, we have labeled five complete two-person meals, including all of the individual activities and group states occurring within them. We report on challenges that arise when attempting to automatically label generalizable features in this context, as well as patterns arising within these categories and the relationships between them. We also describe preliminary attempts to automatically label both individual activities and group states.

A. Our Scenario: Restaurant Dining

We collected our data in a restaurant scenario, where two diners eat a complete meal, beginning as they enter the frame and ending when they both exit. This data was collected from a live public-access webcam in Myrtle Beach, at the RipTydz bar [1]. The overall resolution of the webcam was 1920×1080, but the individual table comprised 450x400 pixels of that space. An annotated image from the webcam of the individual table of interest can be found in Figure 1.

Data Our input is a single-viewpoint RGB video of the scene. The resolution of an individual table is relatively low (400x450), and the majority of poses are seated and in-profile. The scene is highly structured, which allows us to easily identify the static locations of each guest and table elements. Each meal has two in-profile guests sitting in consistent locations relative to the camera, and a standard location where waiters arrive to provide service.

Labeling We have created an action grammar by labeling individual restaurant activities. Each of these activities was chosen to be detectable from externally observable physical actions, i.e. evidenced entirely by the video with no guesswork as to the mental state of the guests. We also created descriptors of collective activities occurring over the course of a meal, referred to as group states.

II. Related Work

Activity Recognition Human activity recognition (HAR) from video is a highly researched problem with many different approaches, with techniques ranging from automatically generating features based on the raw images [15], to generating features and then combining them to produce classifications [11]. However, it is notable that it is often more difficult to capture and correctly classify naturally arising emotional reactions to the world [28], not just those which are performed by actors to create training data [3].

To address this issue, we investigate connecting physically observable individual activities that can be classified from video footage to underlying group states that are more closely linked with emotional drivers such as hunger or impatience. This would avoid relying on extracting these signals from body language itself [2], [14], [31], and better leverage existing methods in HAR.

Our approach also enables proactively identifying opportunities for service, rather than requiring expressions indicating unhappiness [25] or pain [21] before being able to react.

Activities of Daily Living Human activity recognition in restaurants is most similar to activity classification for activities within the home which include cooking and eating. Databases of individual cooking activities [23] have been developed to address this need, as well as studies on first-person activities of daily living among the elderly [30]. These approaches have leveraged static and motion-centric views of the scene [18], as well as silhouette and pose-focused methods [16]. In our work, we examine 3rd person video footage. We also comment on pose as a means of classification that may generalize across individuals given our small number of individuals within the training set.

Human Groupings In human-robot interaction, groupings are recognized as a very common and important phenomenon in characterizing human interactions. Structures such as F-formations [9] describe the manners in which human tend to congregate, and techniques such as that of [12] work to

automatically detect such clusters. These techniques focus on identifying the group as a whole for a robot to approach rather than characterizing within-group dynamics, however, they point to the importance and ubiquity of these kinds of interactions. We use the static structure of the restaurant to identify groupings and relative roles, but we believe that our findings can be combined with these kinds of techniques for dynamic identification of groups and related individuals.

Relationships Between Activities Some approaches to activity recognition have used interdependent elements of the scene as elements of activity recognition. The work of [7] focuses on identifying activities based on the input of a pair of unique and complimentary actors: two hands performing tasks. While this is similar to our approach in that the information about the two actors augments one another, their tight relationship is both task-driven and due to both being controlled by the same actor. Our restaurant scenario supports that a socially-driven, less literal relationship can still provide relevant information for this task.

Similarly, [29] focuses on exploring the predictive power of the relationship between human actors and the objects that they manipulate, particularly in a sports context. Unlike our approach, this relationship is rooted in physical mechanics of the human body and manipulation of tools, a geometric and concrete relationship. The work of [10] exploits the temporal relationship between adjacent frames to augment classification of single frames of video. [27] uses solely the physical locations of teammates to infer puck possession in hockey. These approaches all show that there can be benefits to exploiting contextual relationships within a scene, yet unlike our approach all are grounded in physical interactions of the scene rather than abstract relationships.

III. Restaurant Dataset

This project involved collecting, cleaning, and labeling our restaurant dataset, in addition to pose detection.

A. Terminology

We use the following terms with regards to this problem:

- **Individual Activity** is the observable activity that a single guest is performing. These activities do not require inference about mental state, and can be reliably identified by annotators from video alone. Examples of these kinds of activities are "eating" or "paying bill." Individual activities tend to be on the order of seconds.
- **Group State** refers to the status of the table as a whole. These states require inference on the part of the annotator, and describe overall phases of the meal. Examples of these include states of mutual activity such as "reading menus" as well as inferences about readiness to proceed to the next phase such as "ready to order" and "ready for bill." Group states by definition take into account all individual members of the table, and tend to be on the order of minutes.
- **Target person** is the person whose activity we are trying to classify.
- **Auxiliary person** is not the target of classification, but has the potential to provide additional context for classifying the target person. Multiple auxiliary people are possible, though in our scenario this is always the other group member.
- **Functional Activities** serve a purpose, often progressing the meal or using a tool, such as "paying bill."
- **Social Activities** serve the social aspect of the meal, usually involving interacting with the other person at the table. Examples include "talking," listening behavior "look:partner" and self-entertainment activity "idling."

B. Data Collection

This dataset was collected from an online webcam stream. Wideframe video footage of the entire restaurant was collected, and we screened footage for well-lit two-person meals using the same table and seats with maximal visibility to the fixed camera. Five meals that included recordings of the complete entry and exit of each pair were chosen.

C. Data Labeling

We contend that this naturalistic restaurant data is essential because it provides a high level of ecological validity in this domain. However, due to this data being collected in the wild, determining a ground truth for guest mental states or needs is difficult. We cannot interview guests, and must presume waiter flow is characteristic of reasonable service.

Individual activities were therefore defined to be physical actions which could be observed by an annotator with no additional inference about intent required. For example, "eating" is a label, but "ready to order" requires inferences about guest thinking and therefore is not. Group states require inference, but were consistent across annotators.

This dataset was labelled using video-labeling software Elan [26]. Inter-rater reliability across the two annotators was calculated using the EasyDIAg's technique for comparing the time-event sequential labelings [13]. Labels were consolidated and re-labeled until a kappa score of .85 was achieved. Label "NONE" denotes frames with unclear activities.

Due to the layout of the table within our restaurant, we could distinguish when the target was looking at the other group member, looking at the window, and looking at the direction that the waiter would come from. This specificity is useful for distinguishing a cue of relaxation or boredom ("looking at window") from a potential cue of impatience ("looking for waiter").

D. Pose Availability

To investigate the availability and value of pose data in this domain, we generated pose features for each labeled frame using OpenPose [6]. Each frame contains all detected poses and two labels for the two people in the scene. We then identified which poses belong to each of the two guests by selecting the pose with the most keypoints in the bounding box around the static location of the guest's chair at the table. A visualization of this assignment can be found in Figure 1.

Keypoint	Avail.	Keypoint	Avail.	Keypoint	Avail.
Nose	82.7%	RHip	97.4%	REar	60.7%
Neck	99.0%	RKnee	75.5%	LEar	64.4%
RShoulder	98.4%	RAnkle	67.8%	LBigToe	64.1%
RElbow	82.1%	LHip	97.3%	LSmallToe	62.0%
RWrist	66.7%	LKnee	78.1%	LHeel	64.0%
LShoulder	98.3%	LAnkle	67.8%	RBigToe	60.3%
LElbow	94.0%	REye	47.1%	RSmallToe	56.7%
LWrist	88.6%	LEye	75.4%	RHeel	66.0%
MidHip	98.0%				

TABLE I: *OpenPose joint availability for all five meals.*

IV. Perception Advantages and Challenges

Camera Occlusion Occlusion is a frequent issue within this domain and is highlighted by our single-viewpoint footage. Passers-by can temporarily occlude the guests, and the waiter does consistently. Within the table itself, objects often occluded each other, and the movements of the guests often occlude them as well.

We hold that camera footage is the correct choice for this problem, as it is non-intrusive, and does not require added sensors such as accelerometers or new behaviors by guests. While additional cameras might lessen this effect, even in the best case scenario of first-person camera views, inference and knowledge of hand shape is needed to work around occlusion and aid in identifying pose [20] when interacting with objects. Occlusion will remain an issue in this domain.

Placement of a camera in profile to the participants provides the best vantage point for objects on the table between guests without occlusion, other than waiter visits. This could be provided directly from the perspective of a robot waiter approaching the table, but a static camera communicating with the robot enables the robot to perform other tasks.

Limited Poses For all of our meals, guests are seated except for brief visits out of frame. Our viewpoint does allow for partial observation of leg poses, though this is not guaranteed (ex, a tablecloth or another guest can easily occlude this information). The percent of time each joint is able to be detected across both guests is in table I.

We can see that central points are more reliably found than right and left side joints, and upper body points are more consistently available than lower body. Interestingly, the nose keypoint is less available than other central points. While head angle is important, it can be difficult to detect for both guests from a single camera position.

Static Environmental Features An advantage of the restaurant environment is that the interaction is built around a static area of interest: the table. This allows marking expected locations of chairs, guests, and their associated objects, a fact we leverage for identifying poses as belonging to each guest.

Object Recognition Adjacent domains such as determining interruptibility have found that identifying the objects within the scene helps provide context that indicates when a human is open to being disrupted [4], [5].

Within our dataset, we found that at this resolution only humans, tables, and chairs could be consistently recognized with tools such as YOLO [22]. We found several tools with corresponding activities: *eyeglasses*, *purse*, *wallet*, *napkins*, and *cellphone*. Other key objects found were forks, knives, spoons, plates, glasses, and salt and pepper shakers. Individual utensils are often difficult to identify at this resolution, but can be identified from the pose in which they are held.

We also note that guests often had shared appetizers or sides, meaning that identifying ownership of objects can be somewhat ambiguous in this domain.

	Individual Activity	Average Duration	%
FUNCTIONAL	read:menu	25.0s	8.9%
	use:glasses	4.0s	0.1%
	eating	31.9s	19.2%
	use:napkin	6.6s	2.9%
	read:bill	9.6s	0.3%
	use:wallet	11.1s	0.6%
	use:purse	9.0s	0.2%
	pay:check	20.3s	0.8%
	takeout	106.3s	0.6%
SOCIAL	drinking	7.6s	3.7%
	use:phone	35.9s	9.2%
	talking	12.8s	13.1%
	idle	12.1s	13.9%
	look:partner	9.3s	8.3%
	look:window	8.6s	7.6%
	talk:waiter	10.5s	1.2%
	look:waiter	3.8s	0.7%
MISC	standing	8.9s	0.7%
	away	97.8s	5.8%
	NONE	1.0s	2.2%

TABLE II: *Characterization of individual activities.*

Group State	Average Duration	%
reading-menus	10.9min	18.7%
ready-to-order	0.6min	1.1%
ready-for-food	11.7min	24.0%
eating	16.7min	34.3%
ready-for-cleanup	1.9min	4.0%
ready-for-bill	2.5min	4.3%
paying-bill	0.8min	1.4%
ready-for-final-check	3.2min	5.6%
paying-check	1.2min	2.0%
packing-takeout	2.0min	0.7%
ready-to-leave	2.3min	3.9%

TABLE III: *Characterization of group states observed in our dataset. The overall average meal length was 61 minutes.*

V. Results

We review the collected individual activities and group states, characterizing each. While our sample set of five meals limits the claims of classification effectiveness we can make, our goal is to characterize individual activities for automated classification, and explore the relationship between group states and individual activities.

A. Individual Activities

Statistics on the mean activity length, overall percentage of meal time, and number of meals containing each of our individual activities can be seen in table II, and a histogram of their lengths in 3. We find that overall, these activities

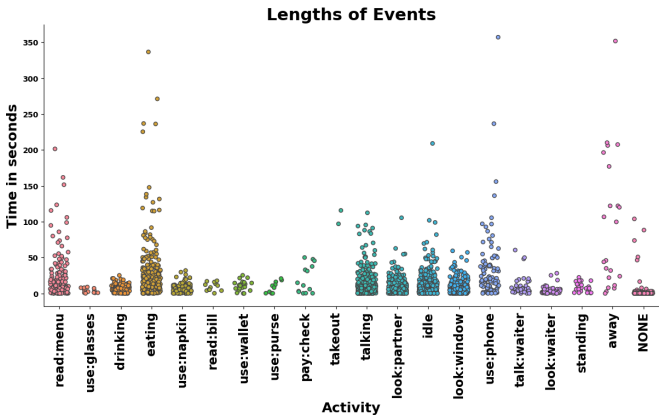

Fig. 3: *Distribution of lengths for each individual activity. Despite the average meal taking over an hour, most activities occur in short bursts on the order of seconds, often interspersed with social activities.*

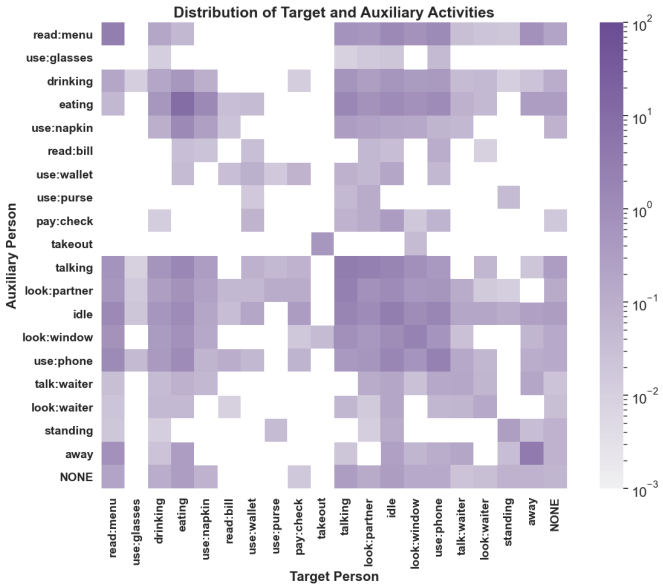

Fig. 4: *Distribution of Target and Auxiliary activities. Labels indicate the percentage of all frames where guests performed this combination of activities, colored along a log scale. The activity observed for an auxiliary person can provide a strong indicator of what activities are most likely for the target person, even without knowledge of group state.*

take place on the order of seconds. For ease of discussion, we have segmented these activities into several categories as seen in II: FUNCTIONAL, SOCIAL, and MISC.

B. Group States

Statistics on the mean activity length, overall percentage of meal time, and number of meals containing each of our individual activities can be seen in Table III. Individual activities emitted within each group state are in Fig. 6.

C. Waiter Visits

Waiter visits are a useful cue for automatically segmenting potential group state transitions, though incomplete. Transitions caused by the waiter can be captured, but these do not include transitions indicating readiness for the waiter to arrive. These "ready" states are almost always transitioned to without waiter action, denoted by a dashed line in Fig. 5.

Explicit Signaling We were unable to detect any instances of individuals directly signaling for the waiter by waving or gesturing in our dataset. This may indicate that waiters usually handle service proactively, and also implies that proactivity would be a required skill for robot waiters.

Labeling of Waiter Events Labeling waiter events is complicated by the video-only format only allowing us to see activities being performed at the table. We cannot hear dialogue. From video, we are able to observe the kinds of objects brought to the table or removed from the table, or use of objects. Waiters are observed to often conduct multiple activities during the same visit, such as both bringing a bill and taking dishes. Noting what activity the waiter is performing over an interval can provide better insight into what underlying needs are being met by their visit. We have noted guesses at waiter actions from observation to add context to Fig. 5, but waiter actions are outside of our focus.

VI. Insights from Data

A. Individual Activity Insights

Other than "away," which is uninterruptible by nature, we find that restaurant activities are highly interlaced. Despite composing 31.9% of the dataset and being the ostensible goal of a meal, "eating" has an average period of 30 seconds, due to being frequently punctuated by other activities.

Knowing that these activities on average take several seconds within meals of roughly 61 minutes, we hypothesize that the length of activities is inversely proportional to how interruptible or multithreadable these activities are. From this perspective, it makes sense that task-oriented activities such as "read:menu," "eating," "pay:check," or preparing "takeout" are among the longest observed. Notably, these activities also correspond to group states of the same name that guests progress through without waiter intervention.

The average 9.6s long "read:bill" activity is an outlier, potentially because this task 1) is during a more social period of the meal 2) that takes a short time to complete compared to the interval and 3) involves review of information rather than decision making or primarily physical action.

Interestingly, self-entertainment tasks such as "use-phone" or "idle" also tend to have a long period (35.9s and 12.1s), potentially indicating that they saturate attention or are function-focused enough to deter interruption.

B. Relationship Between Individual Activities

We began by looking at the relationship between the target and auxiliary individuals. Note that this graph is symmetrical, due to the fact that each of our two diners is the auxiliary to their partner, and vis-versa. This graph is far from evenly distributed; knowledge of one person's activity can rule out categories for the other.

This makes intuitive sense when considering that meals have a series of stages that both members participate in. It

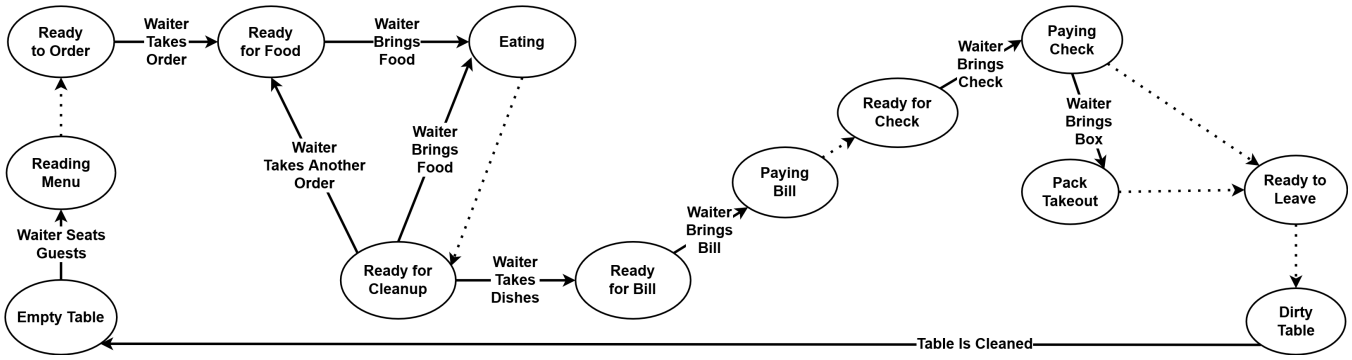

Fig. 5: *Overview of group states and their transitions found in our dataset. All states transitioned to without waiter intervention, denoted by a dashed line, represent states that would particularly benefit from automated detection, as they represent times when the guest's needs have changed and they have become ready for waiter involvement.*

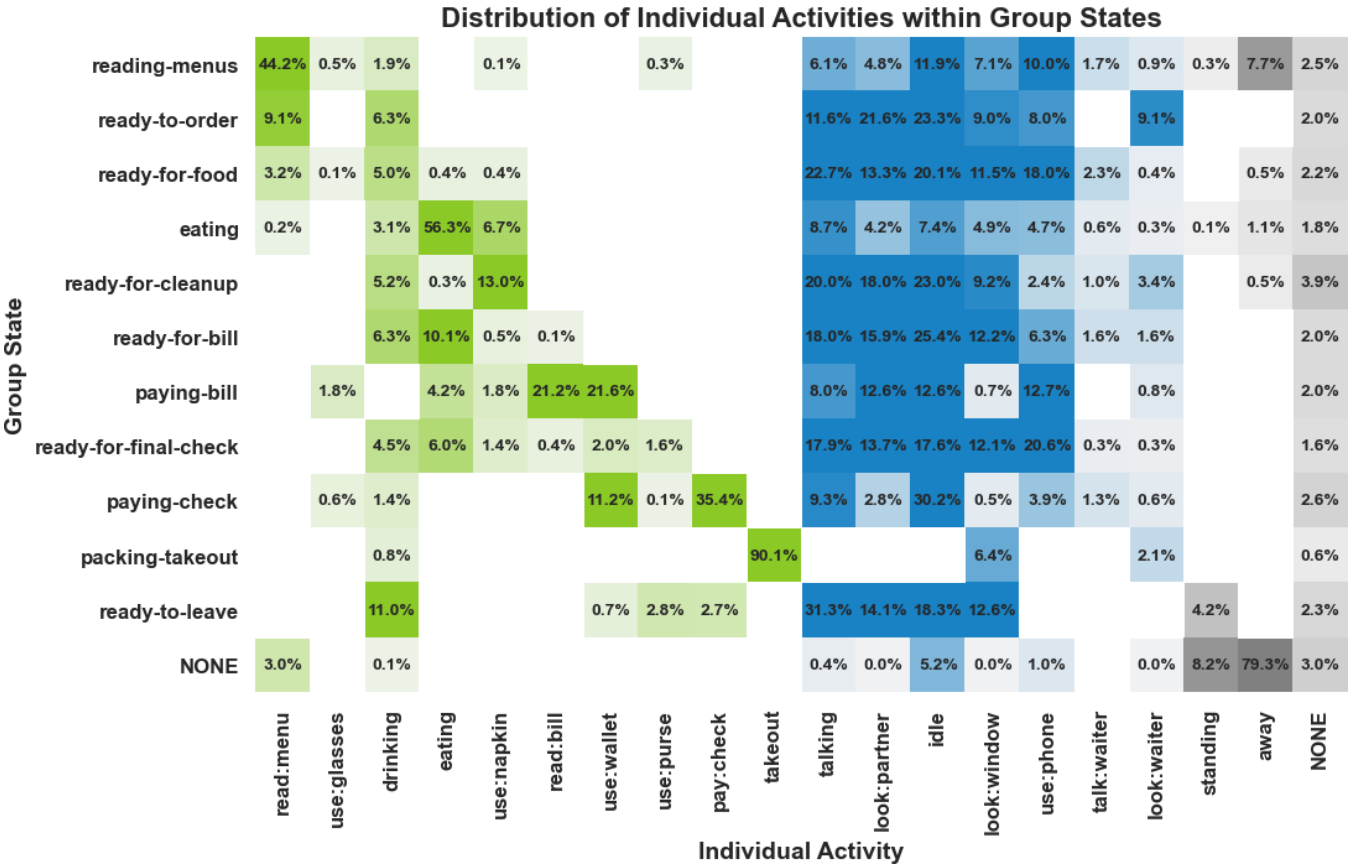

Fig. 6: *For each row representing a group state, we show the percent of time spent emitting each individual activity within each group state. Individual activities are colored green for FUNCTIONAL activities, blue for SOCIAL, and grey for MISC. Each group state has a distinctive profile, and the inclusion of certain activities can provide strong evidence for group state. For example, "use:wallet" strongly supports either "paying-bill" or "paying-check." We also see FUNCTIONAL activities tend to strongly correlate with group states that do not require waiter intervention, as seen by the diagonal clustering across this section in green. SOCIAL activities tend to be spread across all group states, particularly "ready" states.*

is unlikely that one guest will be eating their meal while the other is leaving, for example.

Guests often perform FUNCTIONAL behaviors in unison, as seen from the strong diagonal in the upper left quadrant of Fig. 4. SOCIAL activities often involve combinations of talking and listening behaviors, as seen by the the large square of overlaps seen in the bottom right quadrant. This shows that guests frequently conduct SOCIAL activities in talking and listening combinations such as those observed among "talking," "look:partner," and "idle." We also see relatively many idling activities in unison, notably "idling," "look:window," and "use:phone."

A simple decision tree training on the activity label of the auxiliary guest to predict the target guest's activity can predict the other individual's activity with an accuracy of 35.2%, even without knowledge of the group state as a whole. While still low, this outperforms random chance (4.17%) or majority guess (19.25%) strategies. More sophisticated strategies might blend this information with pose or object information, but even in a state of total occlusion, guessing target activities from auxiliary activities is a useful signal.

These relationships give us reason to believe that adding information about the second individual will potentially improve classification accuracy, and can help to compensate for issues such as occlusion.

C. Necessity of Labeled Group States

We attempted to infer natural underlying cognitive states for the table using a hidden Markov model approach similar to [8], using the Baum-Welch algorithm to build a model of the hidden states that might be producing observations.

We found that this approach has several drawbacks when applied to our domain. This algorithm is non-deterministic, emitting wildly different values for different meals and numbers of states. While we had hoped that these underlying states would have a close relationship to meal phase, waiter visits, or neediness, none of these correlations held true. The significance of these underlying states is unclear, and unlinked from potential waiter actions. An HMM approach of this kind might provide better results for a less linear set of state transitions, such as at a cocktail party with many interactions and low overall state. Our hand-labeled group states provide clarity about what waiter assistance might be required and when, and also have a clear and recurring sequence that can be leveraged for service.

D. Relationship between Individual Activities and Group States

Given labeled and human-understandable group states, we found that the distributions of individual activities do provide an individual "fingerprint" of the group state of group, as seen in Fig. 6.

We can also see that FUNCTIONAL activities correlate strongly with the group states sharing their names: "read:menu," "eating," "pay:check" and "takeout" compose majority actions for the group states sharing their names, and "read:bill" is second only to "use:wallet" within the "paying:bill" group state. Other tool use activities are strong cues for states which rely on them. The exception is "use:phone," which appears to be used across all activities as an idling task, particularly states of "ready-for" or those where the partner may be distracted with a reading task, such as "reading:menus" and "paying:bill." SOCIAL activities take place across all states, potentially reflecting how ubiquitous socialization is in this domain.

While we found that the activities tended to characterize each group state, we also found a degree of "leakage" between states. While, for example, intuitively one might not expect the activity of "reading menu" to occur after the group state proceeded to "ready-to-order", participants would often repeat these behaviors. This may indicate that not all FUNCTIONAL activities are being performed for their direct functional purpose, but for self-entertainment as well.

VII. Additional Discussion

We have underscored the necessity of characterizing the group as a whole in order to solve the problem of providing robotic service at restaurants. Individual activities can be combined to track changes in a table's group state over time, allowing a robot to provide appropriate service when needed. We have also provided evidence for group state's existence, and identified how knowing "ready" states is a desirable goal for activity recognition.

A. Limitations

While we believe the insights from our specific dataset have general applicability, there remain limitations of note in order to apply our work elsewhere.

Generalizability of Annotations Due to the layout of our restaurant, it was easy to annotate glances to the window and glances in the direction of the waiter as separate events. However, not all restaurants will have a window, and not all layouts will enable disambiguation between these two states. We chose to keep this detail rather than just labeling the direction of their glance in hopes of getting better insight into the type of activities performed.

While generalizing methods of detecting individual activities across guests and locations within the restaurant presents challenges, we expect that the trends in how these fuse into group states should hold across tables for a given restaurant.

Human Variability Even within our small dataset, a variety of "wildcard" events such as an extended bathroom break, a visit from a roving entertainer, a child throwing a tantrum on the floor, and extended interactions due to flirtation between guests and the waiter occurred. While our overall insights remain, even across these variations, they do serve to highlight that this domain easily generates unexpected behavior.

Similarly, different groups may have different rates of impatience, behavioral motifs, or underlying needs. This can be seen in some groups having higher rates of behaviors such as "use:napkin", or the fact that stages of service do have variability in length.

B. Future Work

How do we identify relevant social circles? In our scenario, we were able to leverage the static structure of the restaurant and context of two people sitting down to share a table in order to identify the relevant social relationship within the scene. Identifying groups via F-formations [9], [12], shared goals, cohesion when moving together [24], or other indicators of group membership remains future work.

Asymmetrical Roles In our scenario, there is no clear difference in roles between the two tablemates. For other scenarios the matrix in Fig. 4 might not be symmetrical. For larger groups in particular, a leader of the group might be

identified who has a larger impact on group state transitions. Similarly, the person paying the bill may have primary control over bill-related transitions, and potentially earlier payment-related phases such as ordering.

VIII. Contributions

In this paper, we examine the individual activities found in over five two-person restaurant meals with the goal of enabling a robot waiter to determine when guests need its service. We outline the challenges experienced in this domain, as well as useful features for activity classification. We argue that fulfilling the goal of providing proactive robotic restaurant service requires modeling the underlying group state of a table as a whole. We provide support for these group states appearing consistently across meals, and having distinctive profiles for identification. We note relationships of interest between individual activities and group states, and outline how these can be used for improved robot waiter service in the future.

References

[1] Myrtle beach webcams. Available at https://www.myrtlebeach.com/webcams/, accessed 2018.

[2] Meera Alex, Usman Tariq, Fares Al-Shargie, Hasan S Mir, and Hasan Al Nashash. Discrimination of genuine and acted emotional expressions using eeg signal and machine learning. *IEEE Access*, 8:191080–191089, 2020.

[3] Danilo Avola, Luigi Cinque, Alessio Fagioli, Gian Luca Foresti, and Cristiano Massaroni. Deep temporal analysis for non-acted body affect recognition. *IEEE Transactions on Affective Computing*, 2020.

[4] Siddhartha Banerjee and Sonia Chernova. Temporal models for robot classification of human interruptibility. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1350–1359, 2017.

[5] Siddhartha Banerjee, Andrew Silva, and Sonia Chernova. Robot classification of human interruptibility and a study of its effects. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(2):1–35, 2018.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[7] Hapugahage Thilak Chaminda, Vitaly Klyuev, Keitaro Naruse, and Minetada Osano. Recognition of coupling-paired activities in daily life. In *Proceedings of the 2012 Joint International Conference on Human-Centered Computer Environments*, pages 124–130, 2012.

[8] Shih-Yi Chien, Yi-Ling Lin, Pei-Ju Lee, Shuguang Han, Michael Lewis, and Katia Sycara. Attention allocation for human multi-robot control: Cognitive analysis based on behavior data and hidden states. *International Journal of Human-Computer Studies*, 117:30–44, 2018.

[9] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, page 4. Citeseer, 2011.

[10] Georgios Goudelis, Anastasios Tefas, and Ioannis Pitas. Using mutual information to indicate facial poses in video sequences. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–4, 2009.

[11] Harish Haresamudram, David V Anderson, and Thomas Plötz. On the role of features in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 78–88, 2019.

[12] Hooman Hedayati, Daniel Szafir, and Sean Andrist. Recognizing f-formations in the open world. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 558–559. IEEE, 2019.

[13] Henning Holle and Robert Rein. Easydiag: A tool for easy determination of interrater agreement. *Behavior research methods*, 47(3):837–847, 2015.

[14] Mohammed Ehsan Hoque, Daniel J McDuff, and Rosalind W Picard. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, 3(3):323–334, 2012.

[15] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.

[16] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern recognition*, 61:295–308, 2017.

[17] Dongmei Li, Canmian Liu, and Lishan Xie. How do consumers engage with proactive service robots? the roles of interaction orientation and corporate reputation. *International Journal of Contemporary Hospitality Management*, (ahead-of-print), 2022.

[18] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003. IEEE, 2009.

[19] Suzana Marković, Sanja Raspor, and Klaudio Šegarić. Does restaurant performance meet customers' expectations? an assessment of restaurant service quality using a modified dineserv approach. *Tourism and Hospitality Management*, 16(2):181–195, 2010.

[20] Shinya Michibata, Katsufumi Inoue, Michifumi Yoshioka, and Atsushi Hashimoto. Cooking activity recognition in egocentric videos with a hand mask image branch in the multi-stream cnn. In *Proceedings of the 12th Workshop on Multimedia for Cooking and Eating Activities*, pages 1–6, 2020.

[21] Kenneth Prkachin and Zakia Hammal. Automated assessment of pain: Prospects, progress, and a path forward. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 54–57, 2021.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[23] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE, 2012.

[24] Adarsh Jagan Sathyamoorthy, Utsav Patel, Moumita Paul, Nithish K Sanjeev Kumar, Yash Savle, and Dinesh Manocha. Comet: modeling group cohesion for socially compliant robot navigation in crowded scenes. *IEEE Robotics and Automation Letters*, 7(2):1008–1015, 2021.

[25] Naila Sheikh and Mr Kshitij Singhal. A review of human facial expressions recognition methodologies. *Int. J. Sci. Res. Eng. Trends*, 5(2), 2019.

[26] H Sloetjes and P Wittenburg. Elan (version 5.2). nijmegen: Max planck institute for psycholinguistics, 2018.

[27] Moumita Roy Tora, Jianhui Chen, and James J Little. Classification of puck possession events in ice hockey. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 147–154. IEEE, 2017.

[28] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *arXiv preprint arXiv:2203.06935*, 2022.

[29] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1691–1703, 2012.

[30] Kai Zhan, Steven Faux, and Fabio Ramos. Multi-scale conditional random fields for first-person activity recognition. In *2014 IEEE international conference on pervasive computing and communications (PerCom)*, pages 51–59. IEEE, 2014.

[31] Haiyan Zhang, Pengfei Yi, Rui Liu, and Dongsheng Zhou. Emotion recognition from body movements with as-lstm. In *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*, pages 26–32. IEEE, 2021.